(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
7 August 2003 (07.08.2003)

PCT

(10) International Publication Number
WO 03/065350 A1

(51) International Patent Classification[7]: G10L 11/02, 15/24

(21) International Application Number: PCT/IB03/00264

(22) International Filing Date: 29 January 2003 (29.01.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10/058,730 30 January 2002 (30.01.2002) US

(71) Applicant: KONINKLIJKE PHILIPS ELECTRONICS N.V. [NL/NL]; Groenewoudseweg 1, NL-5621 BA Eindhoven (NL).

(72) Inventors: COLMENAREZ, Antonio; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL). KELLNER, Andreas; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL).

(74) Agent: VOLMER, Georg; Internationaal Octrooibureau B.V., Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declaration under Rule 4.17:

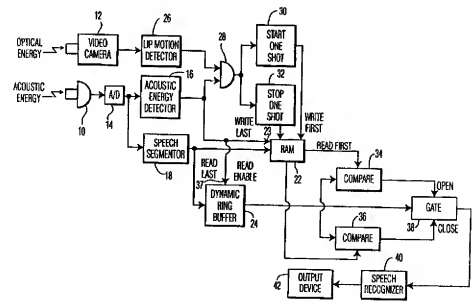
— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii)) for the following designation CN*

Published:

— *with international search report*

[Continued on next page]

(54) Title: AUDIO VISUAL DETECTION OF VOICE ACTIVITY FOR SPEECH RECOGNITION SYSTEM



(57) Abstract: An automatic speech recognizer (40) only responsive to acoustic speech utterances is activated only in response to acoustic energy having a spectrum associated with the speech utterances and at least one facial characteristic associated with the speech utterances. In one embodiment, a speaker must be looking directly into a video camera (12) and the voices and facial characteristics of plural speakers must be matched to enable activation of the automatic speech recognizer.

WO 03/065350 A1

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

AUDIO VISUAL DETECTION OF VOICE ACTIVITY FOR SPEECH RECOGNITION SYSTEM

The present invention relates generally to automatic speech recognition systems and methods and more particularly to an automatic speech recognition system and method wherein an automatic speech recognizer only responsive to acoustic speech utterances is activated only in response to acoustic energy having a spectrum associated with the speech utterances and at least one facial characteristic associated with the speech utterances.

Currently available speech recognition systems determine the beginning and end of utterances by responding to the presence and absence of only acoustic energy having a spectrum associated with the utterances. If a microphone associated with the speech recognition system is in an acoustically noisy environment including, for example, speakers other than the speaker whose voice is to be recognized or activated machinery, including telephones (particularly ringing telephones), the noise limits the system performance. Such speech recognition systems attempt to correlate the acoustic noise with words it has learned for a particular speaker, resulting in the speech recognition system producing an output that is unrelated to any utterance of the speaker whose voice is to be recognized. In addition, the speech recognition system may respond to the acoustic noise in a manner having an adverse effect on its speech learning capabilities.

We are aware that the prior art has considered the problems associated with an acoustically noisy environment by detecting acoustic energy and facial characteristics of a speaker whose voice is to be recognized. For example, Maekawa et al, U.S. patent 5,884,257, and Stork et al, U.S. patent 5,621,858, disclose voice recognition systems that respond to acoustic energy of a speaker, as well as facial characteristics associated with utterances by the speaker. In Maekawa et al., lip movement is detected by a visual system including a light source and light detector. The system includes a speech period detector which derives a speech period signal by detecting the strength and duration of the movement of the speaker's lips. The system also includes a voice recognition system and an overall judgment section which determines the content of an utterance based on the acoustic energy

in the utterance and movement of the lips of the speaker. In Stork et al., lip, nose and chin movement are detected by a video camera. Output signals of a spectrum analyzer responsive to acoustic energy and a position vector generator responsive to the video camera supply signals to a speech classifier trained to recognize a limited set of speech utterances based on the output signals of the spectrum analyzer and position vector generator.

In both Maekawa et al. and Stork et al., complete speech recognition is performed in parallel to image recognition. Consequently, the speech recognition processes of these prior art devices would appear to be somewhat slow and complex, as well as require a significant amount of power, such that the devices do not appear to be particularly well-suited as remote control devices for controlling equipment.

In accordance with one aspect of the present invention, a speech recognition system comprises (1) an acoustic detector for detecting speech utterances of a speaker, (2) a visual detector for detecting at least one facial characteristic associated with speech utterances of the speaker, and (3) a processing arrangement connected to be responsive to the acoustic and visual detectors for deriving a signal. The signal has first and second values respectively indicative of the speaker making and not making speech utterances such that the first value is derived only in response to the acoustic detector detecting a finite, nonzero acoustic response while the visual detector detects at least one facial characteristic associated with speech utterances of the speaker. A speech recognizer for deriving an output indicative of the speech utterances as detected only by the acoustic detector is connected to be responsive to the acoustic detector only while the signal has the first value.

Another aspect of the invention relates to a method of recognizing speech utterances of a speaker with an automatic speech recognizer only responsive to acoustic speech utterances of the speaker. The method comprises: (1) detecting acoustic energy having a spectrum associated with speech utterances, (2) detecting at least one facial characteristic associated with speech utterances of the speaker, and (3) activating the automatic speech recognizer only in response to the detected acoustic energy having a spectrum associated with speech utterances while the at least one facial characteristic associated with speech utterances of the speaker is occurring.

Preferably, activation of the automatic speech recognizer is prevented in response to any of: (1) no acoustic energy having a spectrum associated with speech utterances being detected while no facial characteristic associated with speech utterances of

the speaker is detected, (2) acoustic energy having a spectrum associated with speech utterances being detected while no facial characteristic associated with speech utterances of the speaker is detected, and (3) no acoustic energy having a spectrum associated with speech utterances being detected while at least one facial characteristic associated with speech utterances of the speaker is detected.

In the preferred embodiment, the beginning of each speech utterance is assuredly coupled to the speech recognizer. The beginning of each speech utterance is assuredly coupled to the speech recognizer by: (a) delaying the speech utterance, (b) recognizing the beginning of each speech utterance, and (c) responding to the recognized beginning of each speech utterance to couple the delayed speech utterance associated with the beginning of each speech utterance to the speech recognizer and thereafter sequentially coupling the remaining delayed speech utterances to the speech recognizer. It is assured that no detected acoustic energy is coupled to the speech recognizer upon the completion of a speech utterance. Assurance that no detected acoustic energy is coupled to the speech recognizer upon the completion of a speech utterance is provided by: (a) delaying the acoustic energy associated with the speech utterance, (b) recognizing the completion of each speech utterance, and (c) responding to the recognized completion of each speech utterance to decouple delayed acoustic energy occurring after the completion of each speech utterance from the speech recognizer.

In the preferred apparatus embodiment, the delay is provided by a ring buffer that is effectively indexed so that segmented detected acoustic energy at the beginning of the utterance and segmented detected acoustic energy at the end of the utterance and segmented detected acoustic energy between the beginning and end of the utterance are coupled to the speech recognizer to the exclusion of acoustic energy prior to the beginning of the utterance and acoustic energy subsequent to the end of the utterance.

The processing arrangement in first and second embodiments respectively includes a lip motion and a face recognizer. The face recognizer is preferably arranged for enabling the signal to have the first value only in response to the face of the speaker being at a predetermined orientation relative to the visual detector. The face recognizer also preferably: (1) detects and distinguishes the faces of a plurality of speakers, and (2) enables the signal to have the first value only in response to the speaker having a recognized face.

In the second embodiment, the processing arrangement also includes a speaker identity recognizer for: (1) detecting and distinguishing speech patterns of a plurality of

speakers, and (2) enabling the signal to have the first value only in response to the speaker having a recognized speech pattern.

The above and still further objects, features and advantages of the present invention will become apparent upon consideration of the following detailed description of a specific embodiment thereof, especially when taken in conjunction with the accompanying drawing.

Figure 1 is a block diagram of a preferred embodiment of the speech recognition system in accordance with one embodiment of the present invention; and

Figure 2 is a block diagram of a modified portion of the speech recognition system of Figure 1.

Reference is now made to the Figure 1 of the drawing wherein microphone 10 and video camera 12 are respectively responsive to acoustic energy in a spectrum including utterances of a speaker and optical energy associated with at least one facial characteristic, particularly lip motion, of utterances by the speaker. Microphone 10 and camera 12 respectively derive electrical signals that are replicas of the acoustic and optical energy incident on them in the spectra they are designed to handle.

The electrical output signal of microphone 10 drives analog to digital converter 14 which in turn drives acoustic energy detector circuit 16 and speech segmentor circuit 18 in parallel. Acoustic energy detector 16 derives a bi-level output signal having a true value in response to the digital output signal of converter 14 having a value indicating that acoustic energy above a predetermined threshold is incident on microphone 10. Speech segmentor 18 derives a digital signal that is divided into sequential speech segments, such as phonemes, for utterances of the speaker speaking into microphone 10.

Speech segmentor 18 supplies the sequential speech segments in parallel to random access memory (RAM) 22 and dynamic ring buffer 24. RAM 22 includes an enable input terminal 23 connected to be responsive to the bi-level output signal of acoustic energy detector 16. In response to energy detector 16 deriving a true value, as occurs when microphone 10 is responsive to a speaker making an utterance or ambient noise, RAM 22 is enabled to be responsive to the output of speech segmentor 18. When enabled, sequential memory locations, i.e., addresses, in RAM 22 are loaded with the sequential segments that

segmentor 18 derives by virtue of a data input of the RAM being connected to the segmentor output. This is true regardless of whether the sequential segments are speech utterances or noise. RAM 22 has sufficient capacity to store the sequential speech segments of a typical utterance by the speaker as segmentor 18 is deriving the segments so that the first and last segments of a particular utterance, or noise, are stored at predetermined addresses in the RAM.

Dynamic ring buffer 24 includes a sufficiently large number of stages to store the sequential speech segments segmentor 18 derives for a typical utterance. Thus, buffer 24 effectively continuously records and maintains the last few seconds of acoustic energy supplied to microphone 10. RAM 22 and circuitry associated with it form a processing arrangement that effectively indexes dynamic ring buffer 24 to indicate when the first and last segments of utterances by the speaker who is talking into microphone 10 occur. If the acoustic energy incident on microphone 10 is not associated with an utterance, dynamic ring buffer 24 is not effectively indexed. Buffer 24 is part of a delay arrangement for assuring that (1) the beginning of each speech utterance is coupled to a speech recognizer and (2) upon completion of each utterance the speech recognizer is no longer responsive to a signal representing acoustical energy.

To perform indexing of buffer 24 only in response to utterances by the speaker who is talking into microphone 10, the system illustrated in Figure 1 detects at least one facial characteristic associated with speech utterances of the speaker while acoustic energy is incident on microphone 10. The facial characteristic of the embodiment of Figure 1 is detection of lip motion. To this end, video camera 12 derives a signal indicative of lip motion of the speaker speaking into microphone 10. The lip motion signal that camera 12 derives drives lip motion detector 26 which derives a bi-level signal having a true value while lip motion detector 26 senses that the lips of the speaker are moving and a zero value while lip motion detector 26 senses that the lips of the speaker are not moving.

The bi-level output signals of acoustic energy detector 16 and motion detector 26 drive AND gate 28 which derives a bi-level signal having a true value only while the bi-level output signals of detector 16 and 26 both have true values. Thus, AND gate 28 derives a true value only while microphone 10 and camera 12 are responsive to speech utterances by the speaker; at all other times, the output of AND gate 28 has a zero, i.e., not true, value.

The output signal of AND gate 28 drives one shot circuits 30 and 32 in parallel. One shot 30 derives a short duration pulse in response to the leading edge of the output signal of AND gate 28, i.e., in response to the output of the gate having a transition

from the zero value to the true value. One shot 32 derives a short duration pulse in response to the trailing edge of the output signal of AND gate 28, i.e., in response to the output of the gate having a transition from the true value to the zero value. Hence, one shot circuits 30 and 32 respectively derive short duration pulses only at the beginning and end of a speech utterance. One shot circuits 30 and 32 do not derive any pulses if (1) acoustic energy detector 16 derives a true value while lip motion detector 26 derives a zero value, (2) lip motion detector 26 derives a true value while acoustic energy detector 16 derives a zero value, or (3) neither of detectors 16 nor 26 derives a true value.

The output pulses of one shot circuits 30 and 32 are supplied as write enable signals to first and second predetermined addresses of RAM 22. The first and second addresses are respectively for the first and last speech segments that segmentor 18 derives for a particular utterance. Hence, the first address stores the first speech segment that segmentor 18 derives for a particular utterance, while the second address stores the last speech segment that segmentor derives for that same utterance. RAM 22 is enabled to be responsive to the sequential segments that segmentor 18 derives and the output signals of one shot circuits 30 and 32 by virtue of acoustic energy detector 16 supplying the RAM enable input terminal 23 with a true value during the speech utterance. RAM 22 responds to a transition of the output of acoustic energy detector 16 from a true value to a zero value to read out the contents of the first and second addresses to input terminals of comparison circuits 34 and 36, respectively.

Comparison circuits 34 and 36 are respectively connected to be responsive to the contents of the speech segments stored in the first and second addresses of RAM 22 and the output of dynamic ring buffer 24 to detect the location in the ring buffer of the first and last speech segments of the particular utterance. In particular, upon the completion of a particular speech utterance, RAM 22 supplies (1) one input terminal of comparison circuit 34 with a signal indicative of the speech content of the first speech segment of that utterance and (2) one input terminal of comparison circuit 36 with a signal indicative of the speech content of the last speech segment of that utterance.

While RAM 22 is driving comparison circuits 34 and 36 with the signals indicative of the speech content of the first and last speech segments of the utterance, dynamic ring buffer 24 is enabled by the transition at the trailing edge of the bi-level output of acoustic energy detector 16 to sequentially derive, at a high frequency (i.e., a frequency considerably higher than the frequency at which the segments are transduced by microphone 10), the speech segments it stores. To this end, buffer 24 includes a read out enable input terminal 37 connected to be responsive to the trailing edge transition that detector 16 derives.

While enabled for read out, dynamic ring buffer 24 supplies the sequential speech segments it derives in parallel to second input terminals of comparison circuits 34 and 36.

Comparison circuit 34 derives a pulse only in response to the speech segment that buffer 24 derives being the same as the first segment that RAM 22 supplies to comparison circuit 34. Comparison circuit 36 derives a pulse only in response to the speech segment that buffer 24 derives being the same as the last segment that RAM 22 supplies to comparison circuit 36. Gate 38 has first and second control input terminals respectively connected to be responsive to the output pulses of comparison circuits 34 and 36 and a data input terminal connected to be responsive to the sequential speech segments dynamic ring buffer 24 derives. Gate 38 is constructed so that in response to comparison circuit 34 supplying the first control input terminal of the gate with a pulse, the gate is opened and remains open until it is closed by comparison circuit 36 supplying the second control input terminal of the gate with a pulse.

While gate 38 is open, it passes to automatic speech recognizer 40 the first through the last speech segments dynamic ring buffer 24 supplies to its data input terminal. Automatic speech recognizer 40 can be of any known type that responds only to signals representing acoustic energy and produces an output signal indicative of the speech utterances of the speaker talking into microphone 10 while the speaker is being observed by video camera 12. The output signal of speech recognizer 40 drives output device 42. Examples of output device 42 are a computer character generator for driving a computer display with alphanumeric characters commensurate with the utterances or a machine for performing tasks commensurate with the utterances.

The speech recognition system of Figure 1 can be modified by the arrangement illustrated in Figure 2 so that the speech recognition system will not respond to speech utterances when the speaker is not looking at camera 12 and so that it can respond to speech utterances and the faces of a plurality of speakers. The apparatus illustrated in Figure 2 is connected to respond to the output signal of acoustic energy detector 16, Figure 1, and replaces lip motion detector 26 and AND gate 28.

The apparatus of Figure 2 includes face recognizer 50, connected to be responsive to the output signal of video camera 12, and speaker identity recognizer 52, connected to be responsive to the output signal of acoustic energy detector 16. Face recognizer 50 and speech identity recognizer 52 are connected to other circuit elements and to speech recognizer 40 so that the speech recognizer is activated only when the speaker is facing video camera 12, that is, has a predetermined orientation relative to the video camera.

Hence, if the speaker turns away from and is not looking directly into video camera 12 because the speaker is talking to someone and does not desire to have his/her voice recognized by recognizer 40, recognizer 40 is not activated. Speech recognizer 40 is only activated if the face recognizer 50 and speech recognizer 52 identify the same person. Face recognizer 50 and speech recognizer 52 are trained during at least one training period to recognize the face and speech of more than one person and speech recognizer 40 is activated only if the face and speech are recognized as being for the same person.

To these ends, speaker identity recognizer 52 includes memory 54 having one input connected to be responsive to the speech signal output of analog to digital converter 14 and a second input connected to be responsive to the output of acoustic energy detector 16 so that memory 54 stores short-term utterances of the speaker while detector 16 derives a true value. Upon the completion of the utterance, memory 54 supplies a digital signal indicative of the utterance to one input of comparator 56, having a second input responsive to memory 58 which stores digital signals indicative of the speech patterns of a plurality of speakers who have trained speech recognizer 40.

Comparator 56 derives a true output signal in response to the output signal of speaker memory 54 matching one of the speech patterns that memory 58 stores. Comparator 56 derives a separate true signal for each of the speakers having a speech pattern stored in memory 58. In Figure 2, it is assumed that memory 58 stores speech patterns for first and second different speakers, whereby comparator 56 includes output leads 57 and 59, respectively provided for the first and second speakers. In response to comparator 56 recognizing the speaker as having speech characteristics the same as the speech pattern that memory 58 stores for the first and second speakers, comparator 57 respectively supplies true values to output leads 57 and 59.

Face recognizer 50 includes memory 60 having an input connected to be responsive to the output of video camera 12 so that memory 60 stores one frame of an image being viewed by video camera 12. Upon completion of the frame, memory 60 supplies a digital signal indicative of the frame contents to one input of comparator 62, having a second input responsive to memory 64 which stores digital signals indicative of the facial patterns of each of the plurality of speakers; the facial patterns memory 64 stores are derived while the speakers are looking directly into camera 12, that is, while the faces of the speakers have a predetermined orientation relative to the camera. Comparator 62 derives a true output signal in response to the output signal of memory 60 matching one of the facial patterns that memory 64 stores. Comparator 62 derives a separate true signal for each of the speakers with

facial images stored in memory 64. In the example of Figure 2, memory 64 stores facial images for the first and second speakers, whereby comparator 64 includes output leads 66 and 68, respectively provided for the first and second speakers. In response to comparator 64 recognizing the speaker as having a facial image the same as one of the facial images that memory 60 stores for the first and second speakers, comparator 62 respectively supplies true values to output leads 66 and 68.

During a training period for each of the speakers, each of the speakers recites a predetermined sequence of words, and the speaker is looking directly into video camera 12. At this time, speaker memory 54 is connected to an input of memory 58 to cause the memory 58 to store speech patterns for each of the plurality of speakers who train speech recognizer 40. At the same time, image memory 60 is connected to an input of memory 64, to cause memory 64 to store a facial image for each of the plurality of speakers who train speech recognizer 40. During the training period for each of the speakers, the output of speech segmentor 16 is supplied to the input of speech recognizer 40 to enable the speech recognizer to learn the speech patterns of each of the speakers, in a manner known to those skilled in the art.

The output signals of comparators 56 and 62 on leads 57 and 66 are supplied to inputs of AND gate 70, while the output signals of the comparators on leads 59 and 68 are supplied to inputs of AND gate 72. Hence, AND gate 70 derives a true value only in response to face recognizer 50 and speech identity recognizer 52 both recognizing that a speaker is the first speaker who is looking directly into camera 12. Similarly, AND gate 72 derives a true value only in response to face recognizer 50 and speech identity recognizer 52 both recognizing that a speaker is the second speaker who is looking directly into camera 12. AND gates 70 and 72 derive bi-level signals that are supplied to OR gate 74 which derives a true value in response to either the first or second speakers being identified from the voice and facial characteristics thereof.

The output signal of OR gate 74 drives one shots in the same manner that the output of AND gate 28 drives the one shots. Consequently, the speech signal of the first or second speaker is supplied to speech recognizer 40 in the same manner that the speech signal is supplied to speech recognizer 40 in the embodiment of Figure 1.

To enable speech recognizer 40 of Figure 2 to recognize both speakers, the outputs of AND gates 70 and 72 are supplied to speech recognizer 40. Speech recognizer 40 responds to the outputs of AND gates 70 and 72 to analyze the speech of the correct speaker, in a manner known to those skilled in the art.

While there has been described and illustrated a specific embodiment of the invention, it will be clear that variations in the details of the embodiment specifically illustrated and described may be made without departing from the true spirit and scope of the invention as defined in the appended claims. For example, the discrete circuit elements can be replaced by a programmed computer.

CLAIMS:

1. A speech recognition system comprising an acoustic detector (10) for detecting speech utterances of a speaker; a visual detector (12, 26) for detecting at least one facial characteristic associated with speech utterances of the speaker; a processing arrangement (16, 18, 22, 24, 28, 30, 32, 34, 36, 38) connected to be responsive to the acoustic and visual detectors for deriving a signal having first and second values respectively indicative of the speaker making and not making speech utterances such that the first value is derived in response to the acoustic detector detecting a finite, nonzero acoustic response while the visual detector detects at least one facial characteristic associated with speech utterances of the speaker; and a speech recognizer (40) for deriving an output indicative of the speech utterances as detected only by the acoustic detector, the speech recognizer being connected to be responsive to the acoustic detector in response to the signal having the first value.

2. The speech recognition system of claim 1 wherein the processing arrangement causes the signal to have the second value in response to any of: (a) the acoustic detector not detecting a finite, nonzero acoustic response while the visual detector does not detect speech utterances of the speaker, (b) the acoustic detector detecting a finite, nonzero acoustic response while the visual detector does not detect speech utterances of the speaker, and (c) the acoustic detector not detecting a finite, nonzero acoustic response while the visual detector detects speech utterances of the speaker.

3. The speech recognition system of claim 1 or 2 wherein the processing arrangement includes a delay arrangement (22, 24, 34, 38) for assuring that the beginning of each speech utterance is coupled to the speech recognizer.

4. The speech recognition system of any of claims 1-3 wherein the processing arrangement includes a delay arrangement (22, 24, 36, 38) for assuring that in response to completion of each speech utterance the acoustic detector is decoupled from the speech recognizer.

5. The speech recognition system of claim 3 or 4 wherein the delay arrangement includes a memory element (24) connected to be responsive to the acoustic detector, the memory element including a plurality of stages for storing sequential segments of the output of the acoustic detector, the delay arrangement being such that the contents of the memory element stage storing the beginning of a speech utterance are initially coupled to the speech recognizer.

6. The speech recognition system of claim 5 wherein the memory element includes a ring buffer (24).

7. The speech recognition system of any of claims 1-6 wherein the processing arrangement includes a face recognizer (60, 62, 64) connected to be responsive to the visual detector.

8. The speech recognition system of claim 7 wherein the face recognizer is arranged for enabling the signal to have the first value in response to the face of the speaker being at a predetermined orientation relative to the visual detector.

9. The speech recognition system of claim 7 or 8 wherein the face recognizer is arranged for: (1) detecting and distinguishing the faces of a plurality of speakers, and (2) enabling the signal to have the first value in response to the speaker having a recognized face.

10. The speech recognition system of any of claims 1-9 wherein the processing arrangement includes a speaker identity recognizer (54, 56, 58) connected to be responsive to the acoustic detector, the speaker identity recognizer being arranged for: (1) detecting and distinguishing speech patterns of a plurality of speakers, and (2) enabling the signal to have the first value in response to the speaker having a recognized speech pattern.

11. The speech recognition system of any of claims 1-10 wherein the processing arrangement is arranged for causing the signal to have the first value in response to the speaker having a recognized face matched with a recognized speech pattern of the same speaker.

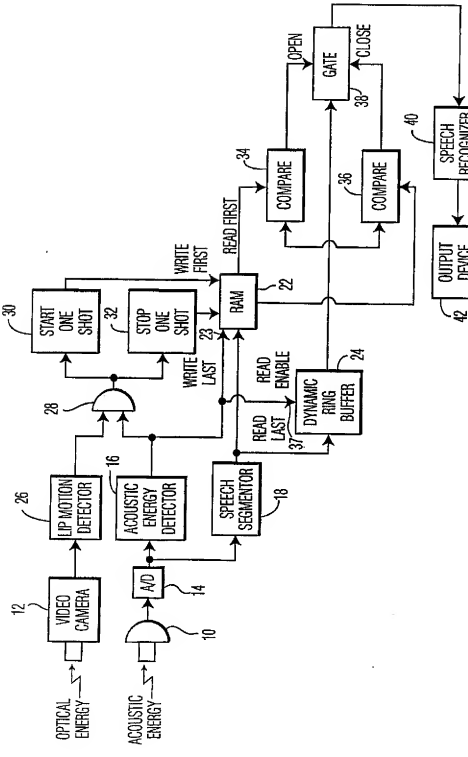

OPTICAL ENERGY
ACOUSTIC ENERGY
12
VIDEO CAMERA
10
A/D
14
26
LIP MOTION DETECTOR
16
ACOUSTIC ENERGY DETECTOR
18
SPEECH SEGMENTOR
28
30
START ONE SHOT
32
STOP ONE SHOT
WRITE FIRST
WRITE LAST
23
RAM
22
READ FIRST
READ LAST
READ ENABLE
37
DYNAMIC RING BUFFER
24
34
COMPARE
36
COMPARE
OPEN
GATE
CLOSE
38
40
SPEECH RECOGNIZER
OUTPUT DEVICE
42

FIG. 1

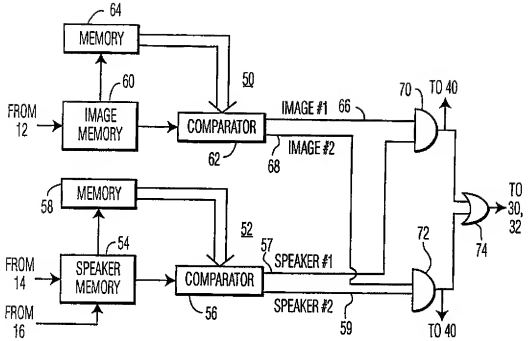

64
MEMORY
60
FROM
12
IMAGE
MEMORY
50
COMPARATOR
IMAGE #1
66
70
TO 40
62
IMAGE #2
68
58
MEMORY
52
54
57
72
TO
30,
32
74
FROM
14
SPEAKER
MEMORY
COMPARATOR
SPEAKER #1
FROM
16
56
SPEAKER #2
59
TO 40

FIG. 2

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L11/02 G10L15/24

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, INSPEC, PAJ, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	PATENT ABSTRACTS OF JAPAN vol. 2000, no. 15, 6 April 2001 (2001-04-06) & JP 2000 338987 A (MITSUBISHI ELECTRIC CORP), 8 December 2000 (2000-12-08) abstract	1,2,7-11
Y	idem	3-6
Y	US 6 216 103 B1 (TANAKA MIYUKI ET AL) 10 April 2001 (2001-04-10) abstract; figure 4 column 1, line 56 - line 66	3-6

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 April 2003	08/05/2003

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Krembel, L

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	GARG A ET AL: "Audio-visual speaker detection using dynamic Bayesian networks" PROCEEDINGS FOURTH IEEE INTERNATIONAL CONFERENCE ON AUTOMATIC FACE AND GESTURE RECOGNITION (CAT. NO. PR00580), PROCEEDINGS OF THE FOURTH INTERNATIONAL CONFERENCE ON AUTOMATIC FACE AND GESTURE RECOGNITION, GRENOBLE, FRANCE, 28-30 MARCH 2000, pages 384-390, XP002238655 2000, Los Alamitos, CA, USA, IEEE Comput. Soc, USA ISBN: 0-7695-0580-5 paragraph '03.2! ---	1,2
A	US 6 219 640 B1 (NETI CHALAPATHY VENKATA ET AL) 17 April 2001 (2001-04-17) figure 1 column 10, line 1 - line 2 column 12, line 9 -column 13 ---	7,9-11
A	CUTLER R ET AL: "Look who's talking: speaker detection using video and audio correlation" 2000 IEEE INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXPO. ICME2000. PROCEEDINGS. LATEST ADVANCES IN THE FAST CHANGING WORLD OF MULTIMEDIA (CAT. NO.00TH8532), PROCEEDINGS OF INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXPO, NEW YORK, NY, USA, 30 JULY-, pages 1589-1592 vol.3, XP002238656 2000, Piscataway, NJ, USA, IEEE, USA ISBN: 0-7803-6536-4 paragraph '02.2! -----	1

Interna Application No
PCT/IB 03/00264

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
JP 2000338987 A	08-12-2000	NONE	
US 6216103 B1	10-04-2001	NONE	
US 6219640 B1	17-04-2001	JP 2001092974 A	06-04-2001

图 14

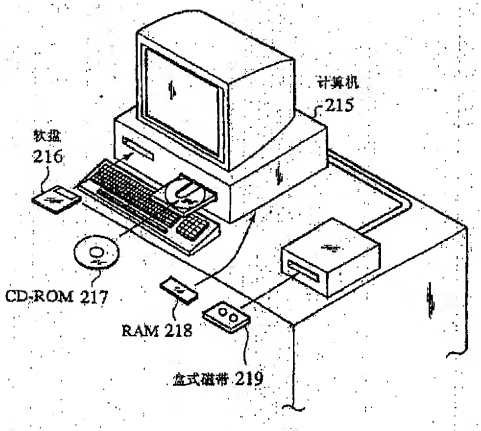


图 13

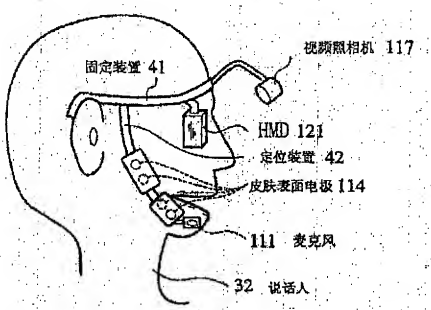


图 12

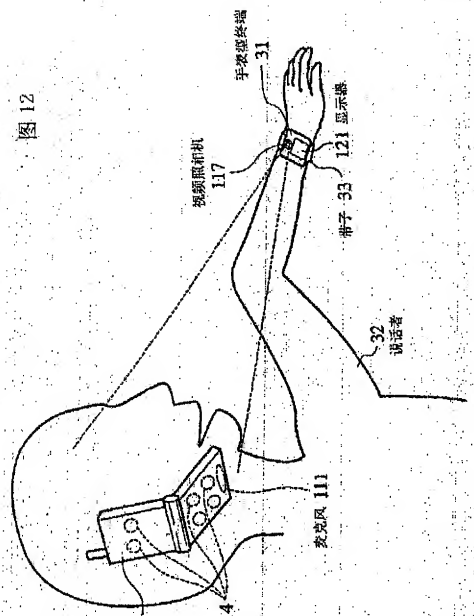
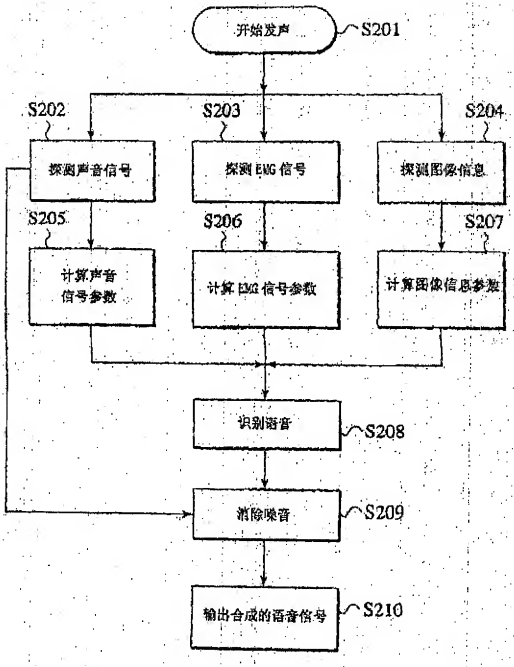

遥控设备 30
皮肤表面电极 114
麦克风 111
视频照相机 117
手表型终端 31
显示器 121
带子 33
32 说话者

图11


开始发声
S201
S202
S203
S204
探测声音信号
探测 EMG 信号
探测图像信息
S205
S206
S207
计算声音信号参数
计算 EMG 信号参数
计算图像信息参数
识别语音
S208
消除噪音
S209
输出合成的语音信号
S210

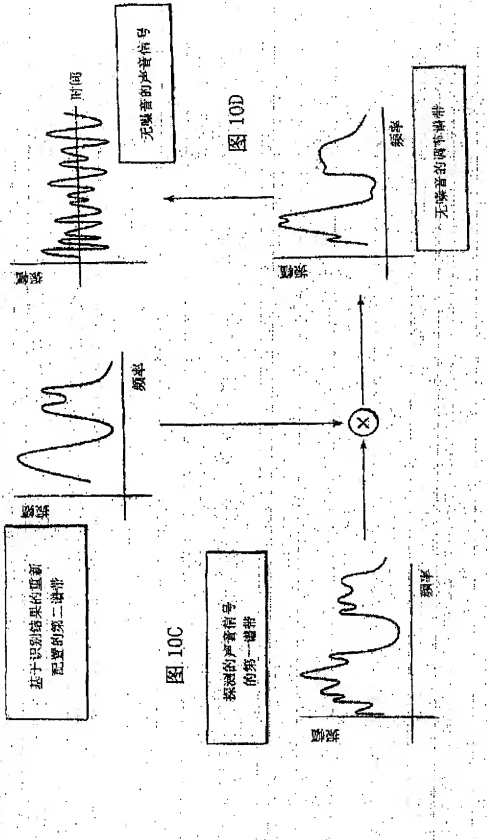


图 10A
幅度
频率
基于识别结果的重新配置的第二滤带
图 10B
幅度
时间
无噪音的声音信号
图 10C
探测的声音信号的第一滤带
幅度
频率
图 10D
幅度
频率
无噪音的调节滤带

图 9

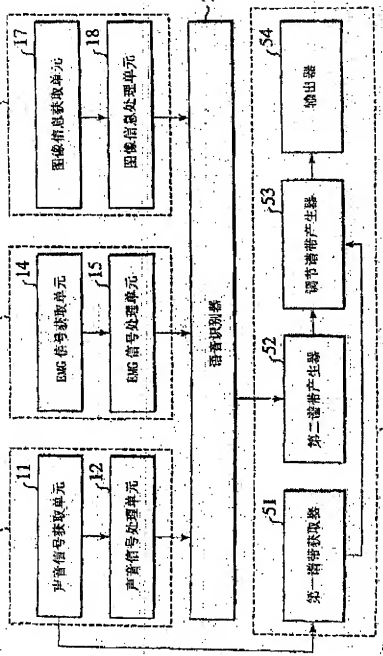

声音信号处理器 10
11
声音信号获取单元
12
声音信号处理单元
EMG信号处理器 13
14
EMG信号获取单元
15
EMG信号处理单元
图像信息处理器 16
17
图像信息获取单元
18
图像信息处理单元
20
语音识别器
51
第一谱带获取器
52
第二谱带产生器
53
调节谱带产生器
54
输出器
语音合成器 55

图 8

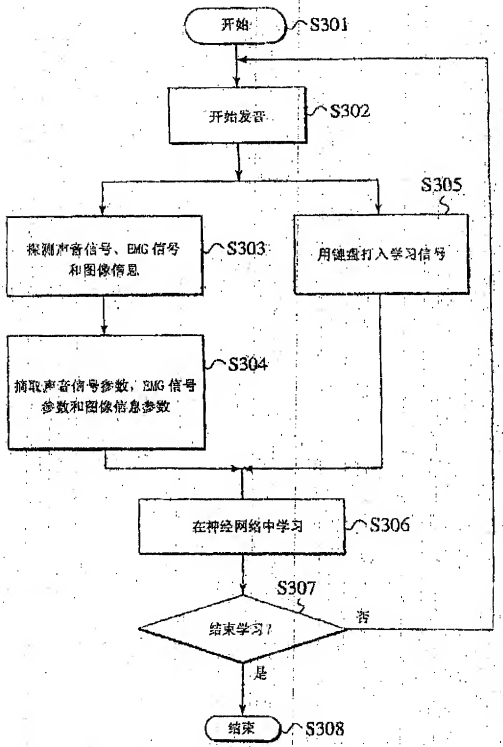
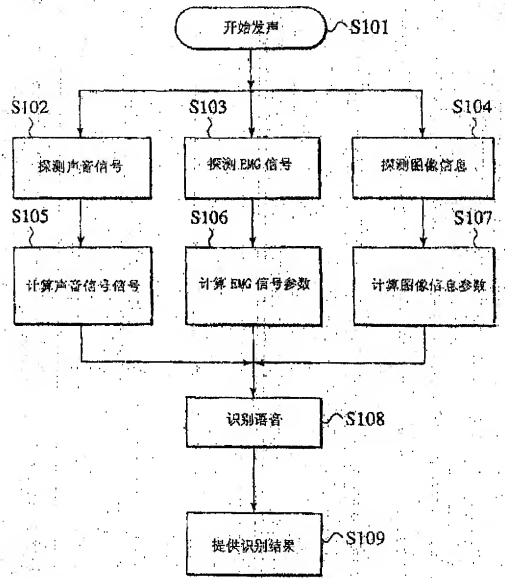

开始
S301
开始发音
S302
S305
探测声音信号、EMG 信号和图像信息
S303
用键盘打入学习信号
摘取声音信号参数，EMG 信号参数和图像信息参数
S304
在神经网络中学习
S306
S307
结束学习？
否
是
结束
S308

图 7


开始发声
S101
S102
S103
S104
探测声音信号
探测 EMG 信号
探测图像信息
S105
S106
S107
计算声音信号信号
计算 EMG 信号参数
计算图像信息参数
识别语音
S108
提供识别结果
S109

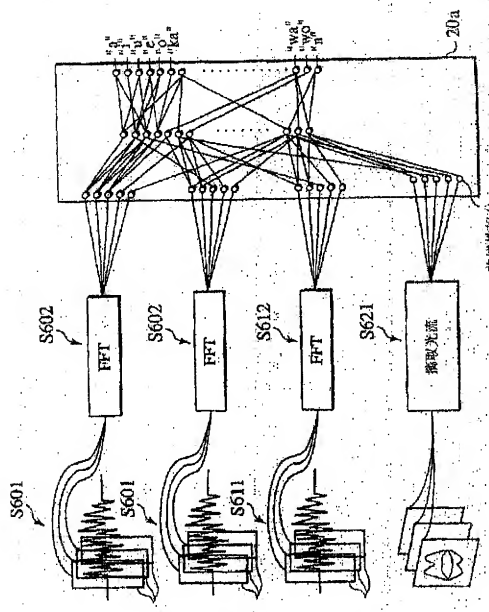

S601
S601
S602
S602
S612
S621
S611
FFT
FFT
FFT
捕获光流
20a
神经网络
非线性组分
EMG 信号 1
时间 - 窗口
EMG 信号 2
时间 - 窗口
声音信号
时间 - 窗口
图像信息

图 6

图 4

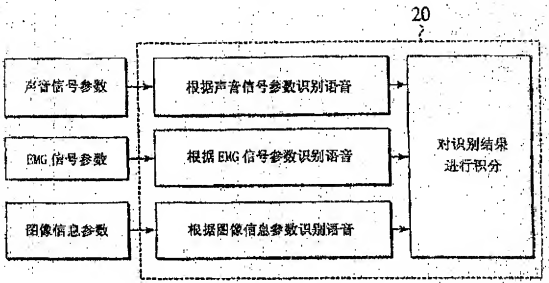
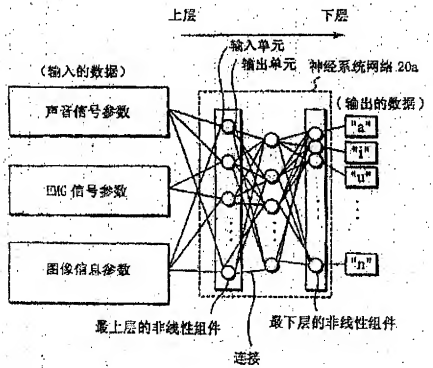
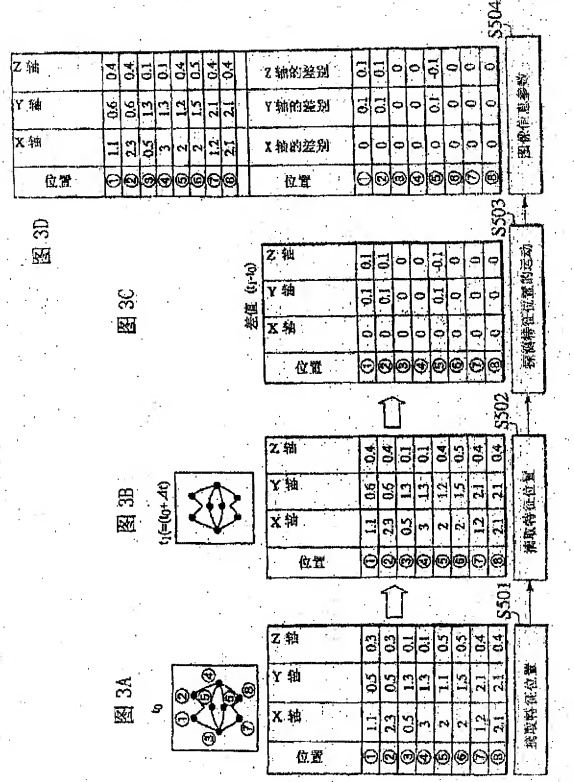

20
声音信号参数
EMG 信号参数
图像信息参数
根据声音信号参数识别语音
根据 EMG 信号参数识别语音
根据图像信息参数识别语音
对识别结果进行积分

图 5


上层
下层
输入单元
输出单元
神经系统网络 20a
(输入的数据)
声音信号参数
EMG 信号参数
图像信息参数
(输出的数据)
"a"
"i"
"u"
"n"
最上层的非线性组件
最下层的非线性组件
连接


S504
S503
S502
S501
量測特徵位置
擷取特徵位置
特徵特徵位置的運動
圖像信息參數
圖 3A
圖 3B
圖 3C
圖 3D
t_0
$t_1=(t_0+\Delta t)$
差值 (t_1-t_0)
位置
X 軸
Y 軸
Z 軸
X 軸的差別
Y 軸的差別
Z 軸的差別

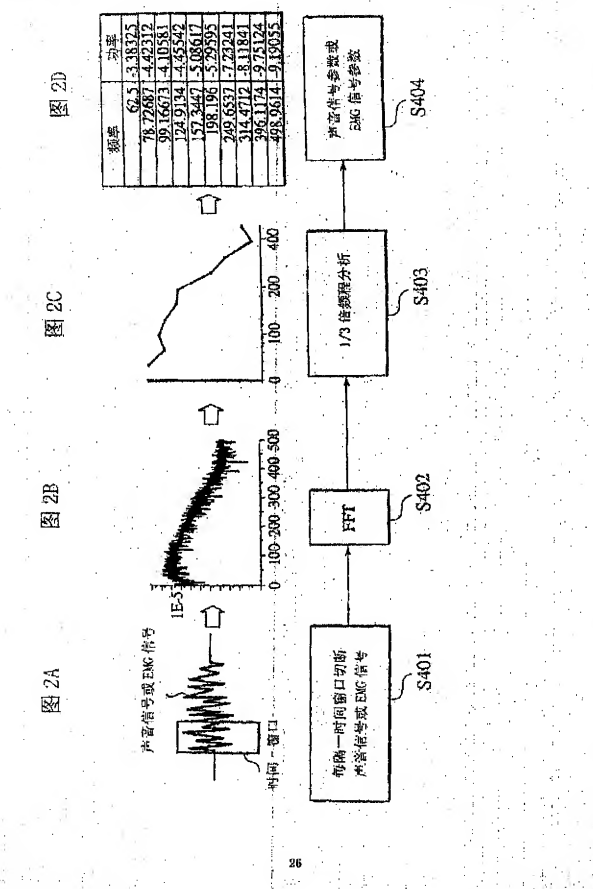


图 2A
图 2B
图 2C
图 2D
声音信号或 EMG 信号
时间 - 窗口
1E-5
0 100 200 300 400 500
0 100 200 400
频率 功率
62.5 -3.38325
78.72687 -4.42312
99.16623 -4.10581
124.9134 -4.45542
157.2447 -5.08617
198.196 -5.59395
249.6537 -7.23241
314.4712 -8.11841
396.1174 -9.75124
498.9614 -9.19055
布响一时间窗口切断声音信号或 EMG 信号
FFT
1/3 倍频程分析
声音信号参数或 EMG 信号参数
S401
S402
S403
S404

图 1

声音信号处理器
10
声音信号获取单元
11
声音信号处理单元
12
EMG 信号处理器
13
EMG 信号获取单元
14
EMG 信号处理单元
15
图像信息处理器
16
图像信息获取单元
17
图像信息处理单元
18
信息积分器
/识别器
19
语音识别器
20
识别结果表供器
21

包含在通信装置 30 中的 IC 芯片或相似的设备上执行用预先确定的程序语言描述的程序来获得。

此外，程序可以记录在存储媒介上，该媒介能够被普通用途的计算机 215 所读取。即，如图 14 所示，程序可以存储在软盘 216、CD-ROM 217、RAM 218、盒式磁带 219 等设备上。通过使用将含有程序的存储媒介插入到计算机 215 或将程序安装到通信装置 30 的内存中等方法可以实现本发明的系统或方法。

与本发明相对应的语音识别系统、方法以及程序在对没有被噪音影响的较低音量的声音信号进行识别时可以保持高的成功率。

与本发明相对应的语音合成系统、方法以及程序能够使用识别的语音信号来合成语音信号，从而使得合成的语音信号更加自然和清晰，并且可以适当地表达说话者的感情等。

线通信或无线通信从语音识别器 20 发送的识别结果进行显示。通信装置 30 可以发送不含噪音的清晰的合成语音信号到手表型的终端 31 中。

在本实施例中，语音识别器 20 被构建在通信装置 30 中，并且构建在手表型的终端 31 中的识别结果提供器 21 显示识别结果。但是，语音识别器 20 也可被安装在手表型的终端 31 中，或其他能与通信装置 30 通信的终端中，该手表型的终端 31 能够识别和合成语音。

识别结果可以从通信装置中输出作为语音信号，可以显示在手表型的终端 31（或通信装置 30）的监视器上，或者可以从另一个能够与通信装置 30 和手表型的终端 31 通信的终端输出。

（根据本发明的第四个实施例的系统）

参考图 13，下面将对用于对根据本实施例的语音识别系统和语音合成系统进行积分的系统进行描述。

如图 13 所示，根据本实施例的系统配置有固定装置 41，该装置作为眼镜形式；作为图像信息获取单元 17 的视频照相机 117，其可被调节以拍摄说话者（声音源）32 的嘴的运动；定位装置 42；作为识别结果提供器的头悬挂显示装置（HMD）121；以及内建于固定装置 41 中的语音识别器 20。固定装置 41 可以悬挂在说话者 52 的头上。

作为 EMG 信号获取单元 14 的皮肤表面电极 114 被配置用来获取说话者 32（声音源）的嘴周围面上的潜在改变，并且作为声音信号获取单元 11 且被配置用来从说话者 32（声音源）的嘴中获取声音信号的麦克风 111 被可调节地固定在说话者 32 的嘴周围。

戴有与根据实施例的系统的说话者 32 能够识别和合成语音，由于使用戴的方式，可以将他/她的双手解放出来。

语音识别器 20 能够内建于固定设备装置 41 中或与固定设备装置 41 进行通信的外部终端中。识别结果可以显示在 HMD（半透明显示设备）中，或作为语音信号从输出设备如扬声器设备中输出。输出设备如扬声器设备能够根据识别结果输出合成的语音信号。

（根据本发明的第五个实施例的系统）

根据上述的实施例的语音识别系统、语音识别方法、语音合成系统或语音合成方法可以通过在普通用途的计算机（例如，个人计算机）215 或

在步骤 S201 中，输出器 54 根据调节谱带输出清晰的合成语音信号。

（根据本发明的第三个实施例的系统）

参考图 12，下面将对整合语音识别系统和语音合成系统的系统进行描述。

如图 12 所示，根据本实施例的系统配置有通信装置 30 以及与之相分离的手表型终端 31。

通讯终端 30 被配置用来添加声音信号处理器 10、EMG 信号处理器 13、语音识别器 20 以及语音合成器 55 到常规的移动终端中。

EMG 信号获取单元 14 包括多个安装的能够与说话者 32 的皮肤接触的皮肤表面电极 114，其被配置用来获得说话者（声音源）32 的嘴周围面上的潜在改变以作为 EMG 信号。声音信号获取单元 11 包括麦克风 111，其被配置用来从说话者（声音源）32 处获取声音信号。麦克风 111 可被配置用来与通信装置 30 进行通信。例如，麦克风 111 被安装在通信装置 30 的表面。麦克风 111 可以为安装在说话者 32 嘴附近的无线麦克风。皮肤表面电极 114 可以被安装在通信装置 30 的表面。

通信终端 30 具有发送基于语音识别器 20 识别的结果而合成的语音信号作为由说话者 32 发出的声音信号的功能。

手表型的终端 31 配置有图像信息处理器 16 和识别结果处理器 21。用于拍摄说话者（声音源）32 的嘴运动图像的视频照相机 117 被安装在手表型的终端 31 上作为图像信息采集单元 17。用于显示识别结果的显示设备 121 被安装在手表型的终端 31 上作为识别结果提供器 21。手表型的终端 31 包括一个用于对其进行固定的带子 33。

对语音识别系统和语音合成系统积分的系统通过安装在通信装置 30 上的 EMG 信号获取单元 14 和声音信号获取单元 11 获得 EMG 信号和声音信号，并且通过安装在手表型的终端 31 上的图像信息获取单元 17 来获得图像信息。

通信装置 30 通过有线通信或无线通信使用手表型的终端 31 对数据进行发送和接收。通信装置 30 和手表型的终端 31 收集并发送信号到构建在通信装置 30 中的语音识别器 20 上，语音识别器 20 根据所收集的信号来识别语音，安装在手表型的终端 31 中的识别结果提供器 21 对通过有

的谱带。如图 10D 所示，更具体地，调节谱带产生器 53 通过用第二谱带（参考图 10A）与第一谱带（参考图 10C）相乘，从而产生没有噪音的调节谱带。

输出器 54 被配置用来根据调节谱带输出合成的语音信号。输出器 54 包括通信装置，其被配置用来发送作为数据的合成的语音信号。如图 10C 所示，更具体地，输出器 54 通过对不含噪音的调节谱带进行傅立叶反向转变（参考图 10D），获得不含噪音的声音信号，并且将获得的声音信号作为合成的语音信号输出。

即是说，根据本实施例的语音合成系统通过将含有噪音的声音信号通过过滤器来获得不含噪音的声音信号，其中过滤器具有由重新配置的谱带所代表的频率特征，并且输出获得的声音信号。

根据本实施例的语音合成系统通过使用各种方法来识别语音，能够将说话者发出的声音信号与周围的噪音从对识别结果进行重新配置得到的信号和声音信号获取单元 11 所探测的信号中分离出来，从而当周围的噪音级别较大时可以输出清晰的合成语音。

因此，根据本实施例的语音合成系统能够在噪音较大或发出的声音信号较小时，输出合成的语音信号，该信号听起来就好像说话者在没有噪音的环境中所发出来的。

根据本实施例的语音合成系统采用了根据第一个实施例的语音识别系统，然而，本发明并不局限于该实施例。根据本实施例的语音合成系统能够根据除声音信号参数以外的参数来识别语音。

参考图 11，下面将对根据本实施例的语音合成系统的操作进行描述。

如图 11 所示，在步骤 S201 到 S208 中，进行与第一个实施例中的识别过程相同的识别过程。

在步骤 S209 中，第一谱带获取器 51 通过声音信号获取单元 11 来获得声音信号的谱带并将其作为第一谱带。第二谱带产生器 52 根据语音识别器 20 识别的结果来产生经过重新配置的声音信号的谱带并将其作为第二谱带。调节谱带产生器 53 根据第一谱带和第二谱带来产生调节后的谱带，在该谱带中噪音（不是说话者所发出的声音信号）已从声音信号获取单元 11 所获得的声音信号中消除。

（根据实施例的语音识别系统的功能及其作用）

本实施例的语音识别系统可以根据从声音信号、EMG 信号以及图像信息计算得到的多个参数来识别语音，从而可以充分地提高抗噪音干扰等能力。

即是说，本实施例的语音识别系统包含三种类型的输入接口（声音信号处理器 10、EMG 信号处理器 13 以及图像信息处理器 16）用于提高抗噪音干扰的能力。当所有的输入接口都不可用时，语音识别系统能够使用可用的输入接口来识别语音，从而提高识别成功率。

因此，本发明能够提供一种语音识别系统，其在周围的噪音级别较大时或当发出的声音信号的音量较小时，能够以足够的级别来识别语音。

（根据本发明的第二个实施例的语音合成系统）

参考图 9 到 11，将对根据本发明的第二个实施例的语音合成系统进行描述。上面所描述的语音识别系统被用于根据本发明的语音合成系统。

如图9所示，根据本发明相的语音合成系统配置有声音信号处理器10、EMG 信号处理器 13、图像信息处理器 16、语音识别器 20 以及语音合成器 55。语音合成器 55 配置有第一谱带获取器 51，第二谱带产生器 52、调节谱带产生器 53 以及输出器 54。

声音信号处理器 10、EMG 信号处理器 13、图像信息处理器 16、语音识别器 20 与第一个实施例中的语音识别系统具有相同的功能。

第一谱带获取器 51 被配置用来获取声音信号的谱带并将其作为第一个谱带，其中声音信号由声音信号获取单元 11 来获取。获取的第一个谱带中含有噪音（参考图 10C）。

第二谱带产生器 52 被配置用来根据语音识别器 20 识别的语音信号（结果）产生经过重新配置的声音信号的谱带并将其作为第二个谱带。如图 10A 所示，更具体地，第二谱带产生器 52 根据从语音识别器 20 识别的结果中摘取的发音音素，例如共振峰频率，来重新配置发音音素的谱带。

调节谱带产生器 53 被配置用来根据第一谱带和第二谱带来产生调节

参考图 7，根据实施例的语音识别系统中进行语音识别的操作。

在步骤 S101 中，说话者开始发声。在步骤 S102 到 S104 中，声音信号获取单元 11、EMG 信号获取单元 14 以及图像信息获取单元 17 分别探测当说话者发声时所产生的声音信号、EMG 信号以及图像信息。

在步骤 S105 到 S107 中，声音信号处理单元 12、EMG 信号处理单元 15 以及图像信息处理单元 18 根据声音信号、EMG 信号以及图像信息分别计算声音信号参数、EMG 信号参数和图像信息参数。

在步骤 S108 中，语音识别器 20 根据计算的参数来识别语音。在步骤 109 中，识别结果提供器 21 提供由语音识别器 20 识别得到的结果。识别结果提供器 21 能够将识别的结果作为语音信号输出或显示识别结果。

其次，参考图 8，为根据本实施例的在语音识别系统中的学习过程的操作。

对于提高识别成功率来说，学习每个说话者的发音特征是很重要的。在实施例中，下面将对使用图 5 中的神经系统网络 20a 进行学习过程的操作进行描述。在不使用神经系统网络 20a 的语音识别方法的情况下，根据本发明的语音识别系统采用了与语音识别方法相关的学习功能。

如图 8 所示，在步骤 S301 和 S302 中，说话者开始发声。在步骤 S305 中，说话者用键盘等输入所说的内容，即是说，当发音时输入学习信号（样品数据）。在步骤 S303 中，声音信号获取单元 11、EMG 信号获取单元 14 以及图像信息获取单元 17 分别探测声音信号、EMG 信号以及图像信息。在步骤 S304 中，声音信号处理单元 12、EMG 信号处理单元 15 以及图像信息处理单元 18 分别摘取声音信号参数、EMG 信号参数和图像信息参数。

在步骤 S306 中，神经系统网络 20a 根据键盘输入的学习信号学习摘取得到的参数。即是说，神经系统网络 20a 通过输入从下到上传送的学习信号（样品数据）来改变指定给非线性组件的加权。

在步骤 S307 中，当识别的错误率低于阈值时，神经系统网络 20a 确定学习过程已经结束。然后操作结束（S308）。

在另一方面，在步骤 S307 中，当神经系统网络 20a 确定学习过程没有完成时，则将重复步骤 S302 到 S306 的操作。

如图 6 所示，被 EMG 信号获取单元 14 探测到的多个 EMG 信号 1，2 在 EMG 处理单元 15（S601）中被放大并且每隔时间一窗口被切断。通过对切断的 EMG 信号进行 FFT 进行谱带的计算。在输入神经系统网络 20 之前，对计算得到的谱带（S602）进行 1/3 倍频程分析，进行 EMG 信号参数的计算。

声音信号获取单元 11 探测到的声音信号被放大并且在声音信号处理单元 12（S611）中每隔时间一窗口进行切断。通过对切断的声音信号进行 FFT 进行谱带的计算。在输入神经系统网络 20 之前，对计算得到的谱带（S612）进行 1/3 倍频程分析，进行声音信号参数的计算。

图像信息处理单元 18 根据图像信息获取单元 17（S621）获取的图像信息来获取说话人嘴周围的特征位置的运动作为光流。作为光流摘取的图像信息参数被输入到神经系统网络 20a 中。

在一连串的时间内拍摄的图像信息中可以摘取嘴周围的各自的特征位置，从而摘取特征位置的运动。也可以将标志放在嘴周围的特征点，并放置参考点，根据探测相对于参考点的特征点的位移，从而摘取特征点的运动。

被输入各种参数的神经系统网络 20a 输出与输入参数相关的音素。

此外，当语音通过如图 4 中的语音识别方法不能够根据任何参数进行识别时，依照本实施例的语音识别器 20 可以被配置用来使用如图 5 中的语音识别方法进行语音识别。通过将图 4 中的语音识别方法识别的结果与图 5 中的语音识别方法识别的结果进行对比或将它们进行积分，语音识别器 20 可以被配置用来对语音进行识别。

识别结果提供器 21 是一种提供（输出）语音识别器 20 识别结果的设备。识别结果提供器 21 能够采用语音产生器将语音识别器 20 识别结果作为语音信号输出到说话人或作为文本信息输出到显示结果的显示器中。识别结果提供器 21 可以包括一个通讯接口，其除了提供结果给说话人外，还传送结果到应用程序中作为数据，该应用程序运行于如个人电脑这样的终端中。

（根据实施例的语音识别系统的操作）

根据实施例的语音识别系统的操作将参考图 7 和图 8 描述如下。首先，

在神经系统网络 20a 中，上层的非线性组件的输出单元被连接到邻近的非线性组件中的下层的非线性组件的输入单元。加权值被指定到该连接或连接的组合。每一个非线性组件根据输入到输入单元的数据以及指定给连接或者组合的加权值来计算从输出单元输出的数据并且确定计算的数据所输出到的连接。

声音信号参数、EMG 信号参数以及图像信息参数被作为输入数据输入到分层网络中的最上层的非线性组件中。识别的语音信号（元音和辅音）被作为输出数据输出到分层语音识别器中的最下层的非线性组件中。语音识别器 20 根据由最下层的非线性组件的输出单元输出的数据来识别语音信号。

通过参考 "Nishikawa and Kitamura," Neural network and control of measure", Asakura Syoten ，18—50 页" 可知，神经系统网络能够采用全连接型的三层神经系统网络。

语音识别器 20 包括学习功能，其被配置用来根据输入的从下向上传送的样品数据来改变指定给非线性组件的加权。

即是说，有必要通过例如反向传播的方法，事先学习神经系统网络 20a 中的加权。

为了学习加权，语音识别器 20 获取根据发出特殊的方式的操作所产生的声音信号参数、EMG 信号参数以及图像信息参数，并且通过使用作为学习信号的特殊的方式来学习加权。

当说话者发音时，EMG 信号比声音信号和图像信息先输入到语音识别系统中，语音识别器 20 通过向神经网络 20a 仅延迟 EMG 信号参数的输入，而不延迟声音信号参数以及图像信息参数的输入，从而使得语音识别器 20 具有同步声音信号、EMG 信号以及图像信息的功能。

接收作为输入数据的各种参数的神经系统网络 20a 输出与输入参数相关的音素。

神经系统网络 20a 采用递归神经系统网络（RNN），其将下一个处理得到的识别结果返回作为输入数据。根据本实施例，语音识别算法除采用神经系统网络外，还可采用各种语音识别算法，例如 Hidden Markov Model（HMM）。

器 20 将在识别上具有最高识别率的识别结果作为最终的结果。

例如，在前面就已经知道的在识别特殊的音素或特殊的说话方式时，根据 EMG 参数进行的语音识别具有较低的成功率，然而，假设特殊的音素或特殊的说话方式被发出，则根据通过非 EMG 信号的参数进行语音识别时，语音识别器 20 忽略根据 EMG 信号参数得到的识别结果，从而可以提高识别成功率。

在基于声音信号参数的语音识别时，当确定周围的噪音级别较大时或发出的声音信号的音量较小时，语音识别器 20 减小基于声音信号参数得到的识别结果对最终结果的影响，并且通过将重点放在基于 EMG 信号参数以及图像信息参数得到的识别结果上来进行语音识别。根据各个参数进行的语音识别可以采用常规的语音识别方法。

基于语音识别器 20 中的声音信号的语音识别可以采用传统的使用各种声音信号的语音识别方法。基于 EMG 信号进行的语音识别可以采用在技术文献“Noboru Sugie et al.,' A speech Employing a Speech Synthesizer Vowel Discrimination from Perioral Muscles Activities and Vowel Production' IEEE transactions on Biomedial Enginnering , 32 卷，第 7 期，485－490 页” 中公开的方法或在 JP－A－181888 等中公开的方法。基于图像信息进行的语音识别可以采用在 JP－A－2001－51963 或 JP－A－2000－206986 等中公开的方法。

如图 4 中所示的语音识别器 20，当声音信号参数、EMG 信号参数以及图像信息参数中的任何参数对于语音识别都没有意义时，例如当周围的噪音级别较大时、当发出的声音信号的音量较小时或当没有探测到 EMG 信号时，语音识别器 20 可以根据有意义的参数来识别语音，从而可在整个语音识别系统中充分地提高对噪音的抗扰性。

参考图 5，下面将对语音识别器 20 的另外一个例子进行具体描述。在图 5 所示的例子中，语音识别器 20 同时根据声音信号参数、EMG 信号参数以及图像信息参数中来识别语音信号。

更加具体的，语音识别器 20 包括一个分层网络（例如，神经系统网络 20a），其中多个包含输入单元和输出单元的非线性组件从上到下被分层地进行定位。

征点的运动（如图 3C 中的 S503）。图像信息处理单元 18 根据计算得到的差值产生图像信息参数（如图 3D 中的 S504）。

对于图像信息处理单元 18 来说，可以使用除在图 3A 到 3D 中的方法以外的其他方法来获取图像信息参数。

图像信息积分器/识别器 19 被配置用来对从声音信号处理器 10、EMG 信号处理器 13 以及图像信息处理器 16 获取的各种信息进行积分和识别。图像信息积分器/识别器 19 配有语音识别器 20 和识别结果提供器 21。

语音识别器 20 通过将声音信号处理器 10 发送的声音信号参数、EMG 信号处理器 13 发送的 EMG 信号参数以及图像信息处理器 16 发送的图像信息参数进行对比和积分，从而进行语音识别的处理器。

语音识别器 20 当周围的噪音级别较小时、当发出的声音信号的音量较大时或当能够根据声音信号参数以足够的级别进行语音识别时，语音识别器 20 能够仅根据声音信号参数来识别语音。

在另一方面，当周围的噪音级别较大时、当发出的声音信号的音量较小时或当不能够根据声音信号参数以足够的级别进行语音识别时，语音识别器 20 不仅能够根据声音信号参数，还能够根据 EMG 信号参数以及图像信息参数来识别语音。

此外，语音识别器 20 能够仅仅根据声音信号参数来识别特殊的音素等，而这种特殊的音素不能够通过使用 EMG 信号参数以及图像信息参数来正确识别，从而可以提高识别的成功率。

参考图 4，下面将对语音识别器 20 的例子进行具体描述。在图 4 所示的例子中，语音识别器 20 根据声音信号参数、EMG 信号参数以及图像信息参数中的每一个来识别语音信号，并将每一个识别的语音信号进行对比，并且根据对比的结果来识别语音信号。

如图 4 所示，更加具体地，语音识别器 20 分别仅根据声音信号参数、EMG 信号参数或图像信息参数来分别识别语音。然后语音识别器 20 根据各个参数对识别的结果进行积分，从而进行语音识别。

当根据各个参数得到的（所有识别结果中的）多个识别结果相互吻合时，语音识别器 20 将这个结果作为最终的识别结果。在另一方面，当根据各个参数得到的（所有识别结果中）没有识别结果相互吻合时，识别

或 EMG 信号被声音信号处理器 12 或 EMG 信号处理器 15 在每时间一窗口时被切断（图 2A 中的 S401）。然后，通过 FFT 由切割信号提取谱带（图 2B 中 S402）。然后，对摘取的谱带进行 1/3 倍频程分析计算出每个频率的功率（图 2C 中 S403）。计算出的与每个频率相关的功率被传输到语音识别器 20 作为语音信号参数或 EMG 信号参数（图 2D 中 S404）。该语音信号参数或 EMG 信号参数被语音识别器 20 识别。

声音信号处理单元 12 或 EMG 信号处理单元 15 也可以通过使用不是在图 2A 到 2D 中的方法来摘取声音信号参数或 EMG 信号参数。

图像信息处理器 16 被配置用于探测当发出声音信号时说话人嘴附近的空间变化。图像信息处理器 16 配置有图像信息获取单元 17 和图像信息处理单元 18。

图像信息获取单元 17 被配置用于通过获取当发出声音信号时说话人嘴附近的空间变化的图像来获取图像信息。图像信息获取单元 17 配置有获取当发出声音信号时说话人嘴附近的空间变化图像的照相机，如视频相机。图像信息获取单元 17 探测嘴附近的运动作为图像信息，并且传送该图像信息到图像信息处理单元 18。

图像信息处理单元 18 被配置用于根据图像信息获取单元 17 获取的图像信息来计算说话人嘴周围的运动参数（图像信息参数）。更具体的，图像信息处理单元 18 用光流摘取嘴周围的运动特征计算图像信息。

参考图 3A 到 3D，下面将对图像信息处理单元 18 进行详细描述。

在说话人嘴附近的特征位置根据时间 t0 时的图像信息进行摘取。（如图 3A 中的 S501）。有可能通过获取标记处的位置作为特征位置，或在拍摄的图像信息中查找特征位置来摘取嘴周围的特征位置。图象信息处理单元 18 能够从图像信息中摘取特征位置并将其作为二维空间位置。图象信息处理单元 18 通过使用多个照相机来获取特征位置并将其作为三维空间位置。

相似地，在经过从 t0 到 t1 这段时间后，在时间 t1 时摘取嘴周围的特征位置（如图 3B 中的 S502）。然后，图像信息处理单元 18 通过计算在时间 t0 时的特征点和在时间 t1 时的特征点之间的差别，计算得到每个特

声音信号获取单元 11 是一种用于从说话人（目标）口中获取声音信号的装置，例如麦克风。声音信号获取单元 11 探测说话人发出的声音信号，并且将获取的声音信号传送到声音信号处理单元 12。

声音信号处理单元 12 被配置用于从声音信号获取单元 11 中获取的声音信号中通过分离谱带包络或微细结构来获取声音信号参数。

声音信号处理单元 12 是一种用于计算声音信号参数的装置，该声音信号参数可以在语音识别器 20 中根据由声音信号获取单元 11 获取的声音信号而被处理。声音信号处理单元 12 每隔一时间一窗口设置时切断声音信号，并且通过常用于语音识别时的分析计算声音信号参数，例如对切断的声音信号进行短时间谱带分析，对数倒频谱分析，最大可能性谱估计方法，协方差方法，PARCOR 分析和 LSP 分析。

EMG 信号处理器 13 被配置用于探测和处理当发出声音信号时说话人嘴附近肌肉的运动。EMG 信号处理器 13 配置有 EMG 信号获取单元 14 和 EMG 信号处理单元 15。

EMG 信号获取单元 14 被配置用于获取（摘取）当发出声音信号时说话人嘴附近肌肉的运动。EMG 信号获取单元 14 探测说话人（目标）嘴附近皮肤表面的可能的变化。也就是说，为了识别嘴附近伴随着发出声音信号的多块肌肉的运动，EMG 信号获取单元 14 通过位于与多块肌肉相关的皮肤表面上的多个电极来探测多个 EMG 信号，并且放大 EMG 信号传输到 EMG 信号处理单元 15。

EMG 信号处理单元 15 被配置用于通过计算由 EMG 获取单元 14 获取的 EMG 信号的功率和分析 EMG 信号的频率来摘取 EMG 信号参数。EMG 信号处理单元 15 是一种根据多个由 EMG 信号获取单元 14 传输的 EMG 信号来计算 EMG 信号参数的装置。更具体的，EMG 信号处理单元 15 在每隔一时间一窗口设置切断 EMG 信号，并且通过计算平均振幅特征，如 RMS（均方根），ARV（平均矫正值）或 IEMG（积分 EMG）来计算 EMG 信号参数。

参考图 2A 到 2D，将对声音信号获取单元 12 和 EMG 信号处理单元 15 进行详细描述。

由声音信号获取单元 11 或 EMG 信号获取单元 14 探测到的声音信号

息的过程的例子。

图 4 为根据本发明的实施例的在语音识别系统中的语音识别器的功能单元图。

图 5 为根据本发明的实施例的在语音识别系统中的语音识别器的功能单元图。

图 6 为在根据本发明的实施例的在语音识别系统中为解释语音识别器的详图。

图 7 为根据本发明的实施例的在语音识别系统中的描述语音识别过程操作的流程图。

图 8 为根据本发明的实施例的在语音识别系统中的描述学习过程操作的流程图。

图 9 为根据本发明的实施例的语音合成系统的功能单元图。

图 10A 到 10D 为在根据本发明的实施例的在语音识别系统中的除去噪音操作的解释图。

图 11 为根据本发明的实施例的在语音系统中描述语音合成过程操作的流程图。

图 12 为根据本发明的实施例的对语音识别系统和语音合成系统一体化系统的完整的配置。

图 13 为根据本发明的实施例的对语音识别系统和语音合成系统一体化的系统的完整配置。

图 14 表示记录了根据本发明的实施例程序的计算机可读记录媒体。

具体实施方式

（根据本发明的第一实施例的语音识别系统的配置）

以下将详细描述根据本发明的第一实施例的语音识别系统的配置。图 1 描述了根据本实施例的语音识别系统的功能单元图。

如图 1 所示，语音识别系统配置有声音信号处理器 10、EMG 信号处理器 13、图像信息处理器 16、信息积分器/识别器 19、语音识别器 20 和识别结果提供器 21。

声音信号处理器 10 被配置用于处理由说话人发出的声音信号。声音信号处理器 10 配置有声音信号获取单元 11 和声音信号处理单元 12。

据声音信号参数、EMG 信号参数以及图像信息参数中的每一个识别语音信号；（D2）对比每个识别的语音信号；以及（D3）根据对比结果识别语音信号。

在本发明的第五个方面的步骤（D）中，语音信号可以同时使用声音信号参数、EMG 信号参数以及图像信息参数来识别。

在本发明的第五个方面，含有输入单元和输出单元的多个非线性组件在分层的网络中从上到下被分层的位置。上层的非线性组件的输出单元连接到邻近的非线性组件中的下层的非线性组件的输入单元。加权值被指定到该连接或连接的组合。每一个非线性组件根据输入到输入单元的数据以及指定给连接或者组合的加权值来计算从输出单元输出的数据并且确定计算的数据所输出到的连接。步骤（D）包括以下步骤：（D11）将声音信号参数、EMG 信号参数以及图像信息参数作为输入数据输入到分层网络中的最上层的非线性组件中；（D12）从分层网络中的最下层的非线性组件的输出单元输出识别的语音信号作为输出数据；并且（D13）根据输出的数据来识别语音信号。

在本发明的第五个方面，计算机可以进行根据输入的从下向上传送的样品数据来改变指定给非线性组件的加权值的步骤。

本发明的第六个方面可归纳为用于在计算机中合成语音信号的程序产品。计算机执行以下的步骤：（A）识别语音信号；（B）获取声音信号；（C）取得获取的声音信号的谱带作为第一谱带；（D）根据语音识别器识别的语音信号来产生声音信号的二次配置谱带，并将其作为第二谱带；（E）根据第一谱带和第二谱带来产生调节后的谱带；以及（F）根据调节后的谱带来输出合成的语音信号。

在本发明的第六个方面中，步骤（F）可以包括发送作为数据的合成的语音信号的步骤。

附图说明

图 1 为根据本发明的实施例的语音识别系统的功能单元图。

图 2A 到 2D 为根据本发明的实施例在语音识别系统中摘取声音信号以及 EMG 信号的过程例子。

图 3A 到 3D 为根据本发明的实施例的在语音识别系统中摘取图像信

别语音信号。

在本发明的第三个方面中，语音信号可以通过在步骤（D）中同时使用声音信号参数、EMG 信号参数以及图像信息参数来识别。

在本发明的第三个方面，含有输入单元和输出单元的多个非线性组件在分层的网络中处于从上到下被分层的位置。上层的非线性组件的输出单元被连接到邻近的非线性组件中的下层的非线性组件的输入单元。加权值被指定到该连接或连接的组合。每一个非线性组件根据输入到输入单元的数据以及指定给连接或者组合的加权值来计算从输出单元输出的数据并且确定计算的数据所输出到的连接。步骤（D）包括以下步骤：（D11）将声音信号参数、EMG 信号参数以及图像信息参数被作为输入数据输入到分层网络中的最上层的非线性组件中；（D12）将识别的语音信号作为输出数据由分层网络中的最下层的非线性组件中输出；并且（D13）根据输出的数据来识别语音信号。

在本发明的第三个方面中，所述方法可以包括根据输入从下层向上层传送的样品数据来改变指定给非线性组件的加权值的步骤。

本发明的第四个方面可归纳为一种语音合成方法，包括以下步骤：（A）识别语音信号；（B）获取声音信号；（C）取得获取的声音信号的谱带作为第一谱带；（D）根据语音识别器识别的语音信号来产生声音信号的二次配置谱带，并将其作为第二谱带；（E）根据第一谱带和第二谱带来产生调节后的谱带；以及（F）根据调节后的谱带来输出合成的语音信号。

在本发明的第四个方面中，步骤（F）可以包括发送作为数据的合成的语音信号的步骤。

本发明的第五个方面可归纳为在计算机中用于识别语音信号的程序产品。计算机执行以下步骤：（A）从对象获取声音信号，并且根据获取的声音信号计算声音信号参数；（B）获取对象的表面的潜在改变作为 EMG 信号，并且根据获取的 EMG 信号计算 EMG 信号参数；（C）取得对象的图像来获取图像信息，并且根据获取的图像信息来计算图像信息参数；（D）根据声音信号参数、EMG 信号参数以及图像信息参数，识别对象发出的语音信号；以及（E）提供语音识别器识别的结果。

在本发明的第五个方面中，步骤（D）可以包括以下步骤：（D1）根

器被安装在主体的表面。

在本发明的第一个方面，系统可以包括一个定位设备以及支撑设备。声音信号处理器可以包括麦克风，其被配置用来从声音源获取声音信号。EMG 信号处理器可以包括电极，其被配置用来获取声音源周围面上的潜在改变以作为 EMG 信号。图像信息处理器可以包括照相机，其被配置用来通过拍摄声音源移动的图像来获取图像信息。定位设备可以固定与声音源接近的麦克风以及电极。支撑设备可以支撑照相机以及定位设备。

在本发明的第一个方面，识别结果提供器可以在半透明的显示设备中显示结果。识别结果提供器被安装在支撑设备中。

本发明的第二个方面可归纳为一种语音合成系统，其包括语音识别器、声音信号获取器、第一谱带获取器、第二谱带产生器、调节谱带产生器以及输出器。

语音识别器被配置用来识别语音信号。声音信号获取器被配置用来获取声音信号。第一谱带获取器被配置用来取得获取的声音信号的谱带来作为第一谱带。第二谱带产生器被配置用来根据语音识别器识别的语音信号来产生声音信号的二次配置谱带，并将其作为第二谱带。调节谱带产生器被配置用来根据第一谱带和第二谱带来产生调节后的谱带。输出器被配置用来根据调节后的谱带来输出合成的语音信号。

在本发明的第二个方面，输出器可以包括通信装置，其被配置用来发送作为数据的合成的语音信号。

本发明的第三个方面可归纳为一种语音识别方法，包括以下步骤：(A) 从对象获取声音信号，并且根据获取的声音信号计算声音信号参数；(B) 获取对象的表面的潜在改变作为 EMG 信号，并且根据获取的 EMG 信号计算 EMG 信号参数；(C) 取得对象的图像来获取图像信息，并且根据获取的图像信息来计算图像信息参数；(D) 根据声音信号参数、EMG 信号参数以及图像信息参数，识别对象发出的语音信号；以及 (E) 提供语音识别器识别的结果。

在本发明的第三个方面中，步骤 (D) 可以包括以下步骤：(D1) 根据声音信号参数、EMG 信号参数以及图像信息参数中的每一个来识别语音信号；(D2) 对比每个识别的语音信号；以及 (D3) 根据对比结果识

并且根据获取的图像信息来计算图像信息参数。语音识别器被配置用来根据声音信号参数、EMG 信号参数以及图像信息参数，识别由对象发出的语音信号。识别结果提供器被配置用来提供语音识别器识别的结果。

在本发明的第一个方面，语音识别器可以根据声音信号参数、EMG 信号参数以及图像信息参数中的每一个来识别语音信号，对比识别的每一个语音信号以及根据对比结果识别语音信号。

在本发明的第一个方面，语音识别器可以同时使用声音信号参数、EMG 信号参数以及图像信息参数来识别语音信号。

在本发明的第一个方面，语音识别器可以包括一个分层网络，在该网络中含有输入单元和输出单元的多个非线性组件被从上到下分层定位。上层的非线性组件的输出单元连接到邻近的非线性组件中的下层的非线性组件的输入单元。加权值被指定给该连接或连接的组合。每一个非线性组件根据输入到输入单元的数据以及指定给连接或者连接的组合的加权值来计算从输出单元输出的数据并且确定计算的数据所输出到的连接。声音信号参数、EMG 信号参数以及图像信息参数被作为输入数据输入到分层网络中的最上层的非线性组件中。识别的语音信号被作为输出数据从分层网络中的最下层的非线性组件中输出。语音识别器根据输出的数据识别语音信号。

在本发明的第一个方面，语音识别器可以包括学习功能，其被配置用来根据输入的从下层向上层传送的样品数据来改变指定给非线性组件的加权值。

在本发明的第一个方面，声音信号处理器可以包括麦克风，其被配置用来从声音源获取声音信号。麦克风被配置用来与通信装置进行通信。EMG 信号处理器可以包括电极，其被配置用来获取声音源周围面上的潜在改变，以作为 EMG 信号。该电极被安装在通信装置的表面。图像信息处理器可以包括照相机，其被配置用来通过拍摄声音源移动的图像来获取图像信息。该照相机被安装在与通信装置分离的终端上。通信装置使用该终端发送和接收数据。

在本发明的第一个方面，终端可包括一个装有照相机的主体，以及固定主体的带子。识别结果提供器可以为用于显示结果的显示器，该显示

Vowel Discrimination from Perioral Muscles Activities and Vowel Production,' IEEE tansactions on Biomedical Engineering, 卷 32，第 7 期，485－490 页"中公开，其中公开了通过将 EMG 信号通过通带过滤器并统计通过的 EMG 信号穿过阈值的次数来区别五个元音字母 "a,i,u,e,o" 的技术。

众所周知，存在通过使用神经系统网络处理 EMG 信号来探测说话者的元音和辅音的方法。此外，使用不只是一个输入渠道而是多个输入渠道输入的信息的多模式接口被提出并已经获取。

在另一方面，传统的语音合成系统存储用于表征说话者的语音信号的数据，并且使用当说话者发声时的数据来合成语音信号。

然而，存在的一个问题是传统的语音探测方法使用从信息而不是从声音信号获取语音信号的技术，因此与使用从声音信号获取语音信号的语音探测方法相比，该技术在识别上具有低的成功率。特别是，很难从嘴内肌肉的运动来识别所发出的辅音。

此外，传统的语音合成系统存在的一个问题在于语音信号是根据表征说话者的语音信号的数据合成的，因此合成的语音信号听起来很生硬，表达不自然，并且不可能确切地表达说话者的感情。

发明内容

综上所述，本发明的一个目的是提供一种语音识别系统和方法，其在没有噪音影响的条件下，识别较低音量的声音信号时具有高的识别率。本发明的另一个目的是提供一种语音合成系统和方法，其使用识别的语音信号来合成语音信号，从而使得合成的语音信号更自然和清晰，并且能够确切地表达说话者的感情。

本发明的第一个方面可归纳为一种语音识别系统，其包括声音信号处理器、肌电图（EMG）信号处理器、图像信息处理器、语音识别器以及识别结果提供器。

声音信号处理器被配置用来从一个对象获取声音信号，并且根据获取的声音信号计算声音信号参数。EMG 信号处理器被配置用来获取对象表面的潜在改变以作为 EMG 信号，并且根据获取的 EMG 信号计算 EMG 信号参数。图像信息处理器被配置用来通过取得对象的图像来获取图像信息，

语音识别系统及方法、语音合成系统及方法及程序产品

技术领域

本发明涉及用于识别语音信号的语音识别系统和方法，根据语音识别进行合成语音信号的语音合成系统和方法以及在其中使用的程序产品。

背景技术

本申请是申请号为 P2002－057818，申请日期为 2002 年 3 月 4 日提出的日本在先专利申请的优先权基础上提出的，该申请的全部内容在此被引入作为参考。

传统的语音探测装置采用语音识别技术通过对发声声音信号中的频率进行分析来对语音信号进行识别和处理。语音识别技术通过使用谱带包络或类似技术获取。

然而，对于传统的语音探测装置来讲，不可能在没有向常规语音探测装置中输入发声的声音信号的条件下来探测语音信号。此外，为了通过使用语音识别技术来获取好的语音探测结果，要求声音信号以一定的音量发出声音。

因此，传统的语音探测装置不能够在需要无声的条件下使用，这些情况例如，在办公室、在图书馆以及在公共机构等地方，当说话者可能会对周围的他/她带来不便时。传统的语音探测装置具有的问题就是在高噪音的条件下，会带来交叉说话的问题并且语音探测功能的性能会降低。

另一方面，已出现了对从除声音信号外的信息获取语音信号的技术的研究。从除声音信号外的信息获取语音信号的技术使得在没有发声的声音信号的条件下获取语音信号成为可能，因此可以解决上述的问题。

根据视频相机输入的图像信息进行图像处理的方法是一种根据嘴唇的视觉信息进行识别语音信号的方法。

此外，还进行了通过处理随着嘴周围（附近）的肌肉运动产生的肌电图（下面称之为 EMG）信号来识别发出的元音类型的技术研究。该研究在"Noboru Sugie 等' 的 A speech Employing a Speech Syntghesizer

（A）从对象获取声音信号，并且根据获取的声音信号计算声音信号参数；

（B）获取对象的表面的潜在改变作为肌电图信号，并且根据获取的肌电图信号计算肌电图信号参数；

（C）取得对象的图像来获取图像信息，并且根据获取的图像信息来计算图像信息参数；

（D）根据声音信号参数、肌电图信号参数以及图像信息参数，识别对象发出的语音信号；以及

（E）提供语音识别器识别的结果。

15．种在计算机中用于合成语音信号的程序产品，其中，计算机执行以下步骤：

（A）识别语音信号；

（B）获取声音信号；

（B）取得获取的声音信号的谱带作为第一谱带；

（D）根据语音识别器识别的语音信号来产生声音信号的二次配置谱带，并将其作为第二谱带；

（E）根据第一谱带和第二谱带来产生调节后的谱带；以及

（F）根据调节后的谱带来输出合成的语音信号。

配置用来获取声音信号的声音信号获取器；

配置用来取得获取的声音信号的谱带作为第一谱带的第一谱带获取器；

配置用来根据语音识别器识别的语音信号产生声音信号的二次配置谱带，并将其作为第二谱带的第二谱带产生器；

配置用来根据第一谱带和第二谱带产生调节后的谱带的调节谱带产生器；以及

配置用来根据调节后的谱带输出合成的语音信号的输出器。

11. 根据权利要求 10 所述的语音合成系统，其中，输出器包括通信装置，其被配置用来传送合成的语音信号作为数据。

12. 一种语音识别方法，包括以下步骤：

（A）从对象获取声音信号，并且根据获取的声音信号计算声音信号参数；

（B）获取对象的表面的潜在改变作为肌电图信号，并且根据获取的肌电图信号计算肌电图信号参数；

（C）取得对象的图像来获取图像信息，并且根据获取的图像信息来计算图像信息参数；

（D）根据声音信号参数、肌电图信号参数以及图像信息参数，识别对象发出的语音信号；以及

（E）提供由语音识别器识别的结果。

13. 一种语音合成方法，包括以下步骤：

（A）识别语音信号；

（B）获取声音信号；

（C）取得获取的声音信号的谱带作为第一谱带；

（D）根据语音识别器识别的语音信号来产生声音信号的二次配置谱带，并将其作为第二谱带；

（E）根据第一谱带和第二谱带来产生调节后的谱带；以及

（F）根据调节后的谱带来输出合成的语音信号。

14. 一种在计算机中用于识别语音信号的程序产品，其中，计算机执行以下步骤：

语音识别器根据输出的数据识别语音信号。

5. 根据权利要求 4 所述的语音识别系统，其中，语音识别器包括学习功能，其被配置用来根据输入的从下层向上层传送的样品数据来改变指定给非线性组件的加权值。

6. 根据权利要求 1 的语音识别系统，其中，

声音信号处理器包括麦克风，其被配置用来从声音源获取声音信号，并且麦克风被配置用来与通信装置进行通信；

肌电图信号处理器包括电极，其被配置用来获取声音源周围表面上的潜在改变，以作为肌电图信号，该电极被安装在通信装置的表面；

图像信息处理器包括照相机，其被配置用来通过拍摄声音源移动的图像来获取图像信息，该照相机被安装在与通信装置分离的终端上，并且

通信装置由该终端发送和接收数据。

7. 根据权利要求 6 所述的语音识别系统，其中，

终端可包括一个装有照相机的主体，以及固定主体的带子，并且识别结果提供器为用于显示结果的显示器，该显示器被安装在主体的表面。

8. 根据权利要求 1 所述的语音识别系统，其中

声音信号处理器包括麦克风，其被配置用来从声音源获取声音信号；

肌电图信号处理器包括电极，其被配置用来获取声音源周围表面上的潜在改变以作为肌电图信号；

图像信息处理器包括照相机，其被配置用来通过拍摄声音源移动的图像来获取图像信息；

定位设备用于固定与声音源接近的麦克风以及电极；

支撑设备可以支撑照相机以及定位设备。

9. 根据权利要求 6 所述的语音识别系统，其中，识别结果提供器可以在半透明的显示设备中显示结果，识别结果提供器被安装在支撑设备中。

10. 一种音合成系统包括：

配置用来识别语音信号的语音识别器；

1. 一种语音识别系统，其包括：

声音信号处理器，其被配置用来从对象获取声音信号，并且根据所获取的声音信号计算声音信号参数；

肌电图信号处理器，其被配置用来获取对象表面的潜在改变以作为肌电图信号，并且根据所获取的肌电图信号计算肌电图信号参数；

图像信息处理器，其被配置用来通过取得对象的图像来获取图像信息，并且根据获取的图像信息来计算图像信息参数；

语音识别器，其被配置用来根据声音信号参数、肌电图信号参数以及图像信息参数识别由对象发出的语音信号；以及

识别结果提供器，其被配置用来提供语音识别器识别的结果。

2. 根据权利要求 1 的语音识别系统，其中，语音识别器根据声音信号参数、肌电图信号参数以及图像信息参数中的每一个来识别语音信号，对比识别的每一个语音信号并且根据对比结果识别语音信号。

3. 根据权利要求 1 得到的语音识别系统，其中，语音识别器同时使用声音信号参数、肌电图信号参数以及图像信息参数来识别语音信号。

4. 根据权利要求 1 的语音识别系统，其中，语音识别器包括一个分层网络，其中含有输入单元和输出单元的多个非线性组件从上到下被分层定位；

上层的非线性组件的输出单元连接到邻近的非线性组件中的下层的非线性组件的输入单元；

加权值被指定给该连接或该连接的组合；

每一个非线性组件根据输入到输入单元的数据以及指定给连接或者组合的加权值来计算从输出单元输出的数据并且确定计算的数据所输出到的连接，

声音信号参数、肌电图信号参数以及图像信息参数被作为输入数据输入到分层网络中的最上层的非线性组件中；

识别的语音信号被作为输出数据从分层网络中的最下层的非线性组件中输出；

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G10L 15/00

[12] 发明专利申请公开说明书

[21] 申请号 03105163.4

[43] 公开日 2003 年 9 月 17 日

[11] 公开号 CN 1442845A

[22] 申请日 2003.3.3 [21] 申请号 03105163.4
[30] 优先权
[32] 2002.3.4 [33] JP [31] 2002-057818
[71] 申请人 株式会社 NTT 都科摩
地址 日本东京都
[72] 发明人 真锅宏幸 平岩明 杉村利明

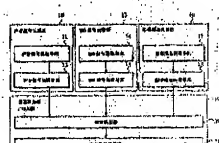
[74] 专利代理机构 北京银龙专利代理有限公司
代理人 徐 川

权利要求书 4 页 说明书 19 页 附图 13 页

[54] 发明名称 语音识别系统及方法、语音合成系统及方法及程序产品

[57] 摘要

本发明涉及用于识别语音信号的语音识别系统和方法，根据语音识别进行合成语音信号的语音合成系统和方法以及在其中使用的程序产品。本发明的语音识别系统包括被配置用来获取声音信号并且根据获取的声音信号计算声音信号参数的声音信号处理器；配置用来获取对象表面的潜在改变以作为肌电图信号，并且根据获取的肌电图信号计算肌电图信号参数的肌电图信号处理器；配置用来通过取得对象的图像来获取图像信息，并且根据获取的图像信息来计算图像信息参数的图像信息处理器；配置用来根据声音信号参数、肌电图信号参数以及图像信息参数，识别由对象发出的语音信号的语音识别器；以及配置用来提供语音识别器识别的结果的识别结果提供器。



ISSN1008-4274